ECSE 6520: Estimation and Detection Theory

Maximum Likelihood Estimation

Class Notes - 9

February 24th, 2014

Contents

1 Introduction

- Maximum likelihood (ML) is one of the most commonly adopted parametric estimation principle in signal processing.
- Unlike other methods, ML usually results in unique estimators and is straightforward to apply to many problems.
- ML estimators have desirable properties.

Definition 1 ***Likelihood Function -*** *For a measurement x we define the* likelihood function *for θ as*

- $L(\theta; x) = f_\theta(x)$
  and the log-likelihood function as
- $\ell(\theta; x) = \log L(\theta; x) = \log f_\theta(x)$.
- Note that sometimes $L(\theta; x)$ is referred to as the likelihood of θ given the measurements x.
- We can view $L(\theta; x)$ as a function of θ parametrically described by the measurements x, whereas $f_\theta(x)$ is a function of x, parametrically described by θ.
- Thus sometimes, the x dependency of the likelihood function is ignored and the notation $L(\theta)$ or $\ell(\theta)$ notations are used.

Definition 2 ***Score Function -*** *If the gradient of the likelihood function*

$$s(\theta, x) = \frac{\partial}{\partial \theta} L(\theta; x)$$

exists, it is called the score function.

2 Maximum Likelihood Principle

ML principle states that the set of model parameters that maximize the apparent probability of a set of observations is the "best set" possible. Maximum likelihood estimator is the implementation of the ML principle.

Definition 3 ***Maximum Likelihood Estimator (MLE) -*** *The estimator $\hat{\theta}$ is called the maximum likelihood estimator if*

$$\begin{aligned} \hat{\theta} &= \arg\max_{\theta\in\Theta} f_\theta(x) \\ &= \arg\max_{\theta\in\Theta} L(\theta; x) \\ &= \arg\max_{\theta\in\Theta} \ell(\theta; x). \end{aligned}$$

- The ML estimator $\hat{\theta}$ is defined as the value of θ which causes the data x to become "most likely," i.e., $\hat{\theta}$ makes it most likely that x was generated from $f(x;\theta)$.
- If the log-likelihood function is differentiable then, the MLE $\hat{\theta}$ satisfies $s(\hat{\theta}) = 0$. We also need to verify that such a solution is a local maximum, not a local minimum or a saddle point. We have to verify that the Hessian, $\nabla^2_\theta L(\theta; x)$, is negative semi-definite at $\hat{\theta}$.
- If several local maximums exists, the MLE is the one with largest likelihood.
- For many members of the exponential family of distributions, it is possible to find a closed form solution for the ML estimator. However, in many cases MLE can not be expressed in closed form. We need to resort to numerical techniques to determine MLE. These techniques include:
  - Newton-Raphson iteration.
  - Expectation Maximization (EM)algorithm.

3 Properties of ML Estimator

- **Property 1.** MLE's are asymptotically unbiased. The proof requires additional technical conditions.

- **Property 2.** MLE's are consistent. The proof requires additional technical conditions.

- **Property 3.** MLE's are asymptotically MVUE in the sense that

$$\lim_{n\to\infty} nVar(\hat{\theta}) = \mathtt{F}(\theta)^{-1}$$

  where F is the Fisher information matrix defined as

$$\mathtt{F}(\theta) = -E[\nabla_\theta^2 \log f_\theta(x)].$$

  For a scalar estimate $\frac{1}{\mathtt{F}(\theta)}$ specifies the fastest possible asymptotic rate of decay of any unbiased estimator's variance. The proof requires additional technical conditions.

- **Property 4.** MLE's are asymptotically Gaussian in the sense

$$\sqrt{n}(\hat{\theta}_n - \theta) \to Z, \quad \textit{in distribution}$$

  where $Z \sim \mathcal{N}(0, \mathtt{F}(\theta)^{-1})$. This means that the cumulative distribution function (cdf) of $\sqrt{n}(\hat{\theta}_n - \theta)$ converges to the (standard normal) cdf of Z. The proof requires additional technical conditions.

- **Property 5.** Unlike many other estimators, e.g. maximum a posteriori (MAP) and MVUE estimators, MLE's are *invariant* to any transformation of the parameters, i.e.,

$$\varphi = g(\theta) \quad \Rightarrow \quad \hat{\varphi} = g(\hat{\theta}).$$

- **Property 6.** The MLE is equivalent to the maximum a posteriori (MAP) estimator (we will cover this topic later) for a uniform prior $f(\theta) = c$.

- **Property 7.** If the MLE is unique, the MLE is a function of the data only through the sufficient statistic.

Efficient Estimators are usually MLE -

Theorem 1 *Assume that the likelihood function $L(\theta; x)$ has at most one local maximum. If $\hat{\theta}$ is efficient, that is $E[\hat{\theta}] = \theta$ and $Cov(\hat{\theta}) = \boldsymbol{F}(\theta)^{-1}$, for all θ, where $\boldsymbol{F}(\theta)$ is the Fisher information matrix, then $\hat{\theta}$ is an MLE.*

4 Application - Estimation of Parameters in Sinusoidal Models

Consider the following model:

$$\begin{aligned} x_t &= s_t + n_t; \qquad t = 0, 1, \ldots, N-1 \\ s_t &= A\cos(\omega t - \phi); \qquad A > 0. \end{aligned}$$

The signal parameters (A, ϕ, ω) are unknown. The noise terms are a sequence of i.i.d. $N[0, \sigma^2]$ random variables, the measurements are a sequence of i.i.d, $N[s_t, \sigma^2]$ random variables. The noise variance is unknown.

The input signal-to-noise-ratio is given by

$$\text{SNR}_{in} = \frac{\frac{1}{N}\sum_{t=0}^{N-1} s_t^2}{\sigma^2} \simeq \frac{A^2}{2\sigma^2} \qquad \text{(large } N\text{)}.$$

The output signal-to-noise-ratio SNR=$N(\text{SNR})_{in}$.

4.1 Joint Density

The joint density function for the random sample $\mathbf{x} = (x_0, x_1, \ldots, x_{N-1})$ is the product density

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = \prod_{t=0}^{N-1} f_{\boldsymbol{\theta}}(x_t) = (2\pi\sigma^2)^{-N/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{t=0}^{N-1}(x_t - s_t)^2\right\}$$

$$s_t = A\cos(\omega t - \phi)$$

$$\boldsymbol{\theta} = (A, \phi, \omega, \sigma^2).$$

The problem is to find the maximum likelihood estimate of $\boldsymbol{\theta}$.

4.2 Likelihood

The log-likelihood of $\boldsymbol{\theta}$ is given by

$$L(\boldsymbol{\theta}, \mathbf{x}) = \ln f_{\boldsymbol{\theta}}(\mathbf{x}).$$

Maximization of likelihood with respect to σ^2 is given by:

$$\frac{\partial}{\partial \sigma^2} L(\boldsymbol{\theta}, \mathbf{x}) = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{t=0}^{N-1} (x_t - s_t)^2 = 0.$$

The maximum likelihood estimate of σ^2 is given by:

$$\hat{\sigma}^2 = \frac{1}{N} v^2$$

where

$$v^2 = \sum_{t=0}^{N-1} (x_t - s_t)^2.$$

When likelihood is jointly maximized with respect to (A, ϕ, ω), then $\hat{\sigma}^2$ becomes the minimum average squared residuals.

Substitute the expression for $\hat{\sigma}^2$ into likelihood:

$$\hat{l}(\boldsymbol{\theta}, \mathbf{x}) = (2\pi v^2/N)^{-N/2} \exp^{-N/2}.$$

4.2.1 Approximation of v^2

The squared residuals may be approximated as follows:

$$\begin{aligned} v^2 &= \sum_{t=0}^{N-1} x_t^2 - 2A \sum_{t=0}^{N-1} x_t \cos(\omega t - \phi) + A^2 \sum_{t=0}^{N-1} \cos^2(\omega t - \phi) \\ &\simeq \sum_{t=0}^{N-1} x_t^2 - 2A \sum_{t=0}^{N-1} \cos(\omega t - \phi) + \frac{A^2 N}{2} \\ &= \sum_{t=0}^{N-1} x_t^2 - 2\mathrm{Re} A e^{j\phi} \sum_{t=0}^{N-1} x_t e^{-j\omega t} + \frac{A^2 N}{2}. \end{aligned}$$

4.2.2 Minimization with respect to ϕ

Differentiate with respect to phase to obtain

$$\frac{\partial}{\partial\phi}\mathrm{Re}\left\{e^{j\phi}\sum_{t=0}^{N-1}x_t e^{-j\omega t}\right\} = \mathrm{Re}\left\{je^{j\phi}\sum_{t=0}^{N-1}x_t e^{-j\omega t}\right\} = 0.$$

The maximum likelihood estimate of ϕ is:

$$\hat{\phi} = -\mathrm{arg}X(\omega)$$

$$X(\omega) = \sum_{t=0}^{N-1}x_t e^{-j\omega t}.$$

where $X(\omega)$ is the discrete-time Fourier transform of the measurements.

4.2.3 Minimization with respect to A

Substitute the solution for ϕ to further compress the likelihood to produce:

$$AN - 2|X(\omega)| = 0$$

which produces the following maximum likelihood estimate of A:

$$\hat{A} = \frac{2}{N}|X(\omega)|.$$

4.2.4 Minimization with respect to ω

Substitute the solution for A to obtain:

$$\begin{aligned} v^2 &= \sum_{t=0}^{N-1}x_t^2 - 2\frac{2}{N}|X(\omega)|^2 + \frac{4}{2N^2}|X(\omega)|^2 N \\ &= \sum_{t=0}^{N-1}x_t^2 - \frac{2}{N}|X(\omega)|^2. \end{aligned} \tag{1}$$

This is minimized by maximizing $|X(\omega)|$. Therefore, the maximum likelihood estimate of ω is

$$\hat{\omega} = \arg\max_{\omega}|X(\omega)|^2.$$

4.3 Summary

$$\begin{aligned}
\hat{\omega} &= \arg\max_{\omega} |X(\omega)|^2 \\
\hat{A} &= \frac{2}{N}|X(\hat{\omega})| \\
\hat{\phi} &= -\arg X(\hat{\omega}) \\
\hat{\sigma}^2 &= \frac{1}{N}\sum_{t=0}^{N-1}\left[x_t - \hat{A}\cos(\hat{\omega}t - \hat{\phi})\right]^2.
\end{aligned}$$

4.4 Cramer-Rao Bounds

Let $\hat{\boldsymbol{\theta}}$ be nay unbiased estimator of $\boldsymbol{\theta}$. The Cramer-Rao Bound says that the error covariance matrix for $\hat{\boldsymbol{\theta}}$ is bounded as

$$\mathbf{C} = E(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T \geq \mathbf{J}^{-1},$$

where $\mathbf{J}$ is the Fisher information matrix:

$$\mathbf{J} = [J_i j]$$

$$J_{ij} = -E\frac{\partial^2}{\partial\theta_i\theta_j}\ln f_{\boldsymbol{\theta}}(\mathbf{x})$$

For this example, $\boldsymbol{\theta} = (A, \phi, \omega, \sigma^2)$.

The natural logarithm of the random variable $f_{\boldsymbol{\theta}}(\mathbf{x})$ is

$$\ln f_{\boldsymbol{\theta}}(\mathbf{x}) = -\frac{N}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=0}^{N-1}(x_t - s_t)^2.$$

From this formula for log likelihood we may differentiate with respect to A, ϕ, ω, and σ^2 to compute the Fisher information matrix. Its inverse is the Cramer-Rao Bound.